공학석사학위논문

# 딥러닝 기반 텍스트 질의응답을 위한 지식 추출 데이터 증강 기법

# Data Augmentation Technique with Knowledge Extraction for Text Question Answering by Deep Neural Networks

2017년 7월

서울대학교 대학원

컴퓨터공학부

조 휘 열

# 딥러닝 기반 텍스트 질의응답을 위한 지식 추출 데이터 증강 기법

# Data Augmentation Technique with Knowledge Extraction for Text Question Answering by Deep Neural Networks

지도교수 장 병 탁

이 논문을 공학석사학위논문으로 제출함.

2017년 7월

서울대학교 대학원

컴퓨터공학부

조 휘 열

조휘열의 석사학위논문을 인준함

2017년 7월

위 원 장 염 헌 영 (인)

부 위 원 장 장 병 탁 (인)

위 원 엄 현 상 (인)

# 초 록

사람과 언어로 의사소통하는 기계를 만드는 것은 튜링테스트를 통한 인공지능 연구자들의 오랜 꿈이다. 그러나 언어의 수많은 예외와 불확실성으로 인해 전통적인 규칙 기반 방식으로는 모델링에 한계가 있었다. 최근 급격히 발전하고 있는 딥러닝 알고리즘을 이용하여 그 한계를 극복한 언어 모델링 연구가 이어지고 있다. 그 중 TextQA는 지문(context)과 질문(question)을 보고 답(answer)을 생성하는 연구로, 언어 모델이 얼마나 언어를 잘 이해하였는가를 테스트하기에 적합하다. 그러나 많은 연구들에도 불구하고 아직 사람의 수준에는 미치지 못하고 있다.

본 논문에서는 TextQA 모델의 성능 향상을 위한 Augmented TextQA를 제안한다. Augmented TextQA는 답을 이용하여 질문을 생성할 수 있는 모델을 만들고, 지문 속에서 추출한 키워드를 입력으로 질문 생성 모델을 통해 질문을 생성한다. 마지막으로, 새로 생성한 데이터를 추가하여 기존 TextQA 모델의 성능 향상을 시도한다. SQuAD(Stanford Question Answering Dataset)와 Seq2Seq 기반 TextQA 모델을 구현하여 실험한 결과, 전반적으로 Augmented TextQA를 사용했을 경우에 성능이 향상되는 것을 확인할 수 있었다. 나아가 Augmented TextQA를 활용하면 외부 자료를 사용하지 않고 도메인 상에서 데이터를 증강시키기 때문에 특정 도메인에서 단어의 표상을 더 적절히 학습할 수 있으며, 학습 데이터에만 국한되지 않은 확장된 TextQA가 가능해지는 효과를 기대할 수 있다.

주요어 : 데이터 증강, 딥러닝, 자연어 처리, 질의응답

학 번 : 2015-22915

# 목 차

# 그림 목차

# I. 서 론

기계에게 인간의 언어를 이해시키고 처리할 수 있게 하는 기술, 자연어 처리는 컴퓨터과학, 인공지능, 언어학 분야에서 꾸준히 연구되어 왔다. 간단하게는 스펠링 체크, 키워드 검색, 동의어 찾기부터 시작해서 정보 추출, 문서 분류를 거쳐 기계 번역, 대화 시스템, 질의응답 시스템에 이르기까지 우리 생활에 항상 존재하는 언어를 처리하는 분야인 만큼 다양한 어플리케이션을 만들어왔다 (Collobert et al., 2011).

하지만 언어는 완벽하지 않다. 동일한 언어일지라도 사용되는 언어의 특성(예: 한국어, 영어, 일본어)과 언어를 사용하는 시간, 장소, 상황, 언어를 사용하는 사람의 가치관과 정서 상태, 언어를 사용하여 의사소통하는 사람간의 암묵적인 동의 등 여러 조건들에 의해 불완전한 문장구조를 가지거나, 중요한 정보가 생략되는 등 수많은 예외와 불확실성을 갖게 된다.

전통적인 자연어 처리 분야에서는 이 불확실성을 처리하기 위해 통계적 언어 모델링 방법을 사용해왔다. 단어 예측 문제를 예로 들면, $n$번째 단어 을 예측하기 위해서 앞서 나왔던 $w_1, w_2, \ , w_{n-1}$들을 이용하여 주어진 corpus, 혹은 단어 사전에 있는 각 단어의 확률 $(w_n | w_1, w_2, \ , w_{n-1})$을 계산하고자 하였다. 그러나 계산 과정에서 모든 단어 간의 dependency를 고려하는 것이 불가능에 가깝기 때문에 independent assumption이나 markov assumption을 가정한 naive bayes 모델 (Chai et al., 2002), N-gram 모델 (Brown et al., 1992), hidden markov 모델 (Eddy, 1996) 등이 주로 활용되었다.

딥러닝의 등장으로 다양한 분야에서 높은 성능 향상이 이루어지자, 자

연어 처리에서도 딥러닝 알고리즘을 적극 활용하기 시작했다. 딥러닝의 주요 특징은 양질의 feature들을 이용하여 각 단어에 대한 표상(representation)을 자동적으로 학습하는 것인데, 이러한 특징을 자연어에 적용하여 불확실성까지 최대한 모델링 하고자 하였다. 딥러닝에서 단어의 표상은 다차원의 실수 벡터로 표현하는 distributed representation hypothesis를 따른다 (Rieger, 1991). 기존 자연어 처리에서는 해당 단어가 몇 번 등장했는지에 대한 frequency 기반 표상을 사용했던 반면, 딥러닝에서는 해당 단어의 사전적, 문맥적 의미가 벡터의 여러 차원에 반영될 것이라는 가정 하에 각 단어를 다차원의 실수 벡터 형태로 표현하여 해당 벡터를 학습한다. 그렇게 학습한 단어의 표상, word vector를 이용하여 딥러닝은 여러 자연어 처리 문제를 해결하고 있다.

다양한 자연어 처리 문제 중 TextQA는 텍스트를 잘 이해했는지를 판단할 수 있는 좋은 기준이 된다. TextQA는 지문(context), 질문(question), 답(answer)으로 구성되어 있는 데이터를 이용하여, 지문과 지문에서 답을 찾을 수 있는 질문을 보고 답을 생성하는 문제이다. 따라서 TextQA에서 뛰어난 성능을 내기 위해서는 지문을 통해 올바른 표상을 만들고, 질문을 잘 이해한 뒤, 구축한 단어 사전 속에서 정답을 올바르게 인출하는 것이 중요하다. 좋은 표상을 만들기 위해서 Dynamic Memory Network (Kumar et al., 2016), r-net (Wang et al., 2017)과 같은 자연어를 이해하기 위해 복잡한 모델들을 만드는 연구들이 이루어지고 있으며, 부족한 데이터를 극복하기 위해 유사한 다른 데이터에서 학습한 표상을 사용하는 전이 학습 (transfer learning) (Pan & Yang, 2010), 학습할 때 word vector들이 좋은 시작점에서부터 학습할 수 있도록 값을 초기화해주는 word2vec (Le & Mikolov, 2014), GloVe (Pennington et al., 2014) 와 같은 pretrained word vector

representation 방법이 사용된다. 그 중 전이 학습과 pretrained word vector representation을 활용한 방법들은 주어진 데이터외의 다른 데이터들을 사용한다. 그러나 단어에는 어감이 존재하여 같은 단어라도 상황에 따라 다른 의미로 쓰일 수 있다. 그러므로 무분별하게 다양한 도메인에서 많은 데이터를 수집하여 학습하는 것은 좋은 표상을 생성하여 성능을 향상시키는데 큰 도움이 되지 않을 수 있다.

본 논문에서는 외부 데이터의 활용 없이 한 도메인 내에서 TextQA 모델의 성능 향상을 위한 Augmented TextQA를 제안한다. Augmented TextQA는 답을 이용하여 질문을 생성할 수 있는 모델을 만들고, 지문 속에서 추출한 키워드를 입력으로한 질문 생성 모델을 통해 새로운 질문을 생성한다. 마지막으로, 증강된 데이터를 추가하여 기존 TextQA 모델의 성능 향상을 시도한다.

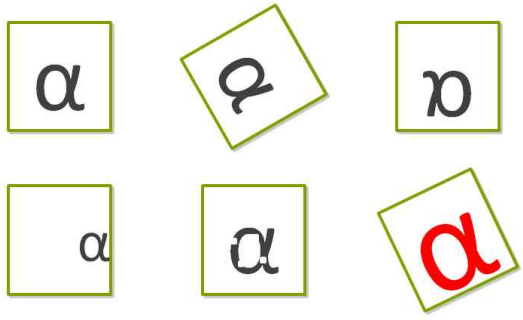
먼저 2.1장을 통해 이미지에서 많이 사용되고 있는 데이터 증강 기법과 그 아이디어를 간략하게 설명한 후, 2.2장, 2.3장에서 TextQA, 질문 생성 모델의 기본이 되는 Seq2Seq 모델 그리고 키워드 추출에 활용되는 Stanford CoreNLP 도구에 대해 각각 기술할 것이다. 3장에서는 실험에 활용한 Stanford Question Answering Dataset과 전체적인 실험 설계, 각 파트에 들어갈 후보 신경망 모델 구조를 기술한다. 4.1장에서는 자연어 데이터의 전처리 과정과 전처리 과정에 따른 성능 차이를 보인다. 그리고 4.2장에서는 설계한 모델들을 이용하여 TextQA 성능을 보이고 4.3장에서 입력과 출력이었던 답과 질문을 반대로 하여 모델이 질문을 생성하도록 학습한다. 4.4장에서는 질문 생성에서 성능이 가장 뛰어났던 모델과 Stanford CoreNLP를 가지고 지문 속 키워드를 새로운 입력으로 하여 데이터를 생성하며, 마지막 4.5장에서는 증강된 데이터를 활용한 Augmented TextQA의 성능을 보고한다. 5장과 6장에서는 기존 TextQA

와 Augmented TextQA의 성능을 비교하고, Augmented TextQA의 장점과 단점, 그리고 한계점을 논할 것이다.

# II. 이론적 배경

## 2.1. 데이터 증강 기법

데이터 증강 기법이란, 인위적으로 데이터의 양을 증가시키는 기술이다 (Wan et al., 2014). 주로 이미지 데이터에 사용되어 왔는데, 이미지의 레이블을 변경하지 않고 이미지의 픽셀을 변화시켜 변형된 데이터를 이용해서 추가적인 학습을 한다. 이미지 데이터에 사용되는 주요 데이터 증강 방법으로는 Horizontal/Vertical Flips, Random Crops/Scales, Color Jittering 등이 있다. Flip은 이미지의 정보를 손상시키지 않는 선에서 반전 시키는 것이고, Crops/Scale은 이미지를 알아볼 수 있는 범위 내에서 특정 부분을 잘라 내거나 이미지의 크기를 조절하는 것이며 Color Jittering은 이미지의 색상을 변경하는 것이다 [그림 1]. 특정 물체의 사진을 좌우 반전 시킬지라도 우리에게는 같은 물체지만, 픽셀 단위로 이미지를 인식하는 컴퓨터에게는 다른 데이터로 인식된다. 따라서 컴퓨터는 변형된 이미지를 다른 이미지로 인식하여 새로운 데이터로 학습할 수 있다.

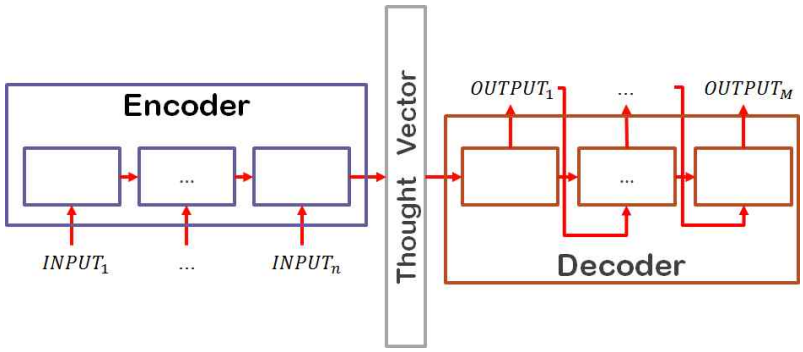


**[그림 1]** 증강된 이미지 데이터 예시

반면 자연어 처리에서는 문장 속 단어의 순서에 따라 문장의 의미가 달라질 수 있으며 단어의 형태 또는 철자에 따라서도 의미가 변화하기 때문에 이미지에서 사용한 데이터 증강 기법을 그대로 활용할 수 없다. 이에 Zhang & LeCun (2015)은 자연어 처리에서 데이터 증강 기법을 활용할 수 있는 방법으로 유의어 사전을 이용한 문장 속 단어의 치환이나 사람이 직접 보고 의역 (paraphrasing)하는 방법을 제시했다. 하지만 이 방법은 방대한 유의어 사전을 구축해야한다는 단점이 있으며 구축한다고 할지라도 유의어 사전 속에서 치환하기에 가장 적절한 단어의 기준을 정하기가 모호하다는 문제점이 있다. 그 다음으로 사람을 통해 일일이 글을 의역하기에는 시간과 비용이 많이 든다는 문제점이 있다. 따라서 우리는 딥러닝 알고리즘 중 문장 생성에 용이한 Seq2Seq (Sutskever et al., 2014) 모델을 사용하여 모델이 직접 새로 학습할 QA 데이터를 자동 생성하도록 하였다.

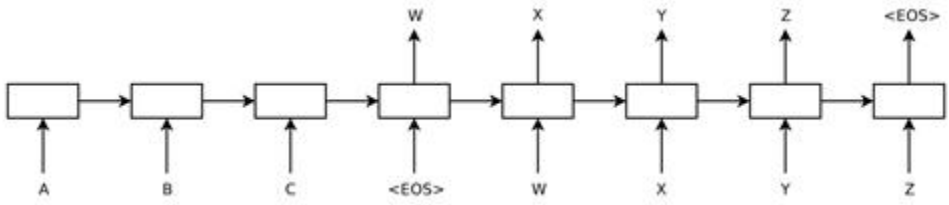
## 2.2. Seq2Seq 모델

Seq2Seq 모델 (Sutskever et al., 2014) 은 인코더-디코더 모델의 하나로, 연속적인 속성을 가진 데이터를 입력으로 받아, 연속적인 속성을 가진 데이터를 출력하는 모델이다. 인코더-디코더 모델은 입력을 처리하는 인코더 부분과 처리된 결과를 풀어 결과를 생성하는 디코더 부분으로 이루어져있다 [그림 2].




**[그림 2]** 인코더-디코더 모델

일반적인 Seq2Seq 모델은 두 개의 RNN (Recurrent Neural Networks)으로 구성되어 있는데, 첫 번째 RNN은 한 문장을 단어 단위로 쪼개어 하나의 벡터로 인코딩하는 작업을 한다. 그리고 두 번째 RNN은 첫 번째 RNN에서 인코딩한 벡터를 하나의 단어씩 디코딩하여 최종적으로 하나의 문장을 만든다. 감독학습 방법을 이용하여 입력 문장을 넣었을 때 정해진 출력 문장이 나오도록 학습 한다.



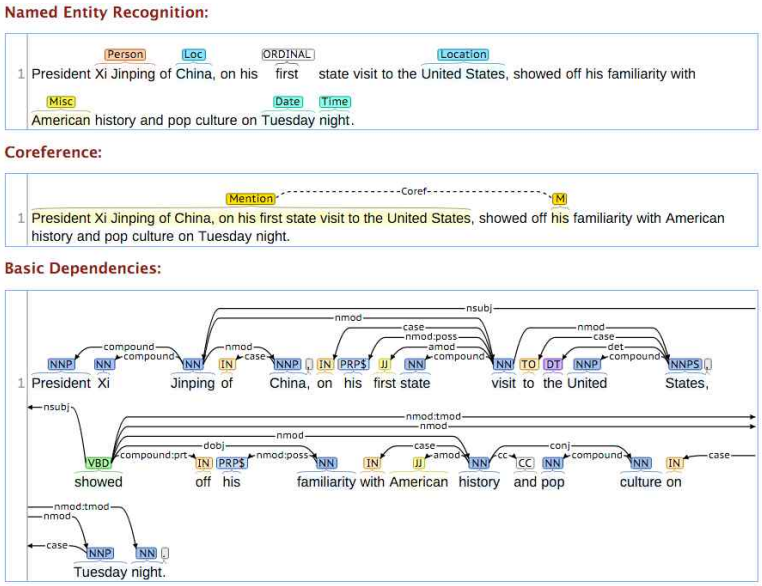

**[그림 3]** Seq2Seq 모델 예시 (출처: Sutskever et al., 2014)

위 [그림 3]는 "ABC"라는 입력이 들어왔을 때 "WXYZ"라는 출력을 내는 Seq2Seq 모델의 예시이다. "ABC"의 문장은 각각 'A', 'B', 'C'의 단어 혹은 토큰 단위로 쪼개져 모델의 입력 값으로 들어간다. "ABC"의 문장이 전부 들어간 뒤에는 문장의 끝을 뜻하는 <EOS> 심볼이 입력 값으로 들어가면서 모델은 한 단어씩 디코딩을 시작한다. 가장 먼저 <EOS> 심볼이 입력되면 정답 문장의 첫 번째 단어인 W가 출력되고 출력된 W를 다시 디코더에 입력시켜 다음 단어인 'X'를 출력한다. 계속해서 이전 출력을 다음 입력으로 넣어 디코딩을 하며 정답 문장과의 차이가 감소하는 방향으로 학습을 한다.

Seq2Seq는 RNN 두 개를 연결하는 간단한 방법으로 순차적인 데이터를 효과적으로 처리할 수 있는 구조를 제공해주었다. 그리고 기존 자연어 처리에서 가변적인 길이를 처리하기 위해 수행하였던 많은 작업들을 단순화시키며, 입력을 넣고 출력을 바로 얻는 end-to-end의 편리함도 누릴 수 있게 해주었다. Seq2Seq는 문장을 생성하는 자연어 처리에 적극 활용되어 기계 번역 (Luong et al., 2015), 음성 인식 (Bahdanau et al., 2016), 이미지 캡션 생성 (Mostafazadeh et al., 2017), TextQA (Dong & Lapata, 2016) 등에 널리 활용되고 있다.

## 2.3. Stanford CoreNLP

Stanford CoreNLP (Manning et al., 2014)는 영국 Stanford NLP 그룹에서 배포한 자연어 처리 도구 소프트웨어이다. 대표적으로 개체명 인식(Named Entity Recognition), 품사 태깅 (Part-Of-Speech), 그 밖에 coreference resolution, sentiment analysis, bootstrapped pattern learning 등 다양한 작업을 수행할 수 있다 [그림 4]. 특히 Named Entity Recognizer를 활용하여 미리 정의해 둔 장소, 시간, 단위 등에 해당하는 단어를 인식하고 POS Tagger를 이용하여 각 단어의 문법 정보를 활용하여 키워드 추출에 활용할 수 있다. 뛰어난 성능으로 키워드 추출뿐만 아니라 다른 자연어 처리 어플리케이션의 전처리에 많이 사용되고 있다.

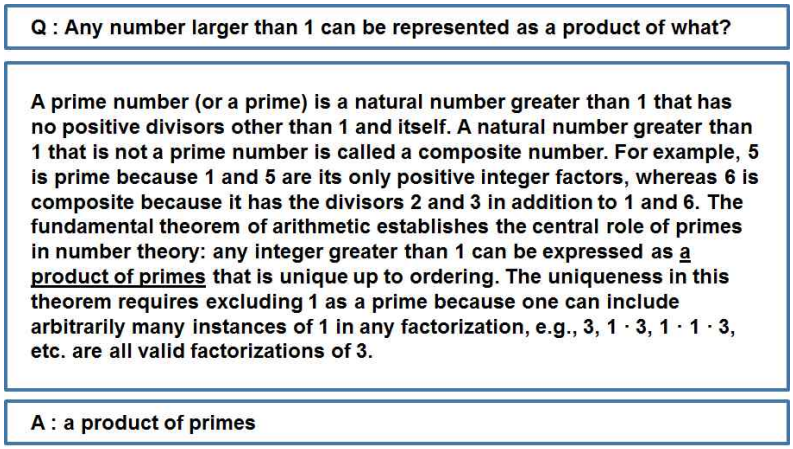



**[그림 4]** Stanford CoreNLP 활용 예시
(출처: https://stanfordnlp.github.io/CoreNLP/)

# III. 연구 방법

## 3.1. 데이터

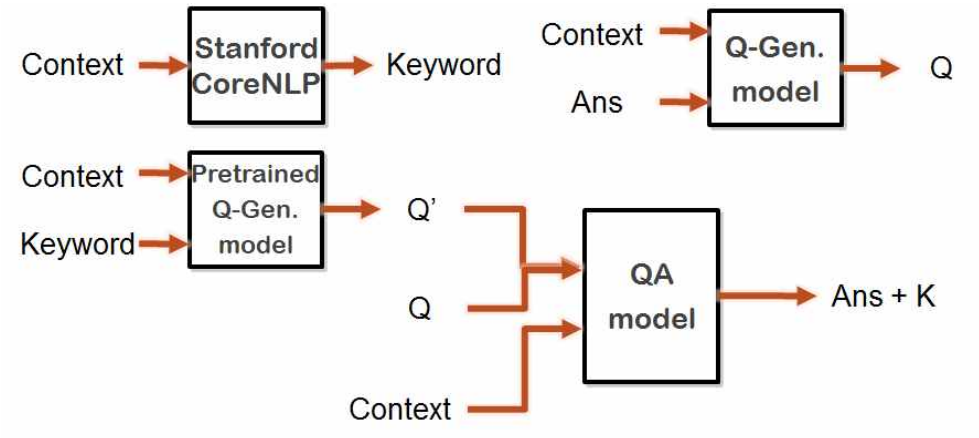
실험에는 SQuAD (Stanford Question Answering Dataset)을 사용했다. SQuAD (Rajpurkar et al., 2016)는 영국 Stanford 대학에서 공개한 Reading Comprehension 데이터셋으로 500개 이상의 위키피디아 문서에 대해 총 100,000개 이상의 QA pair를 만들었다. 기존의 데이터셋들은 데이터의 양(Quantity)과 질(Quality) 사이의 tradeoff에서 한쪽 측면에 치우쳐있는 반면, SQuAD는 비교적 적절한 양의 데이터와 그 질을 만족했다. SQuAD 데이터의 주된 특징으로는 답(answer)이 항상 지문(context)에 있다는 것이다. 따라서 이 데이터에서는 TextQA 문제를 답 문장의 첫 번째 단어의 index를 찾는 문제로 바꿔 해결할 수도 있다. [그림 5]는 Prime Number에 대한 지문, 그리고 QA 예시이다. 실제 데이터에는 지문의 제목, 답의 index 정보가 추가로 있으며 답은 하나씩만 제공된다.




**[그림 5]** SQuAD 데이터 예시

## 3.2. 설 계

### 3.2.1 전체 구조

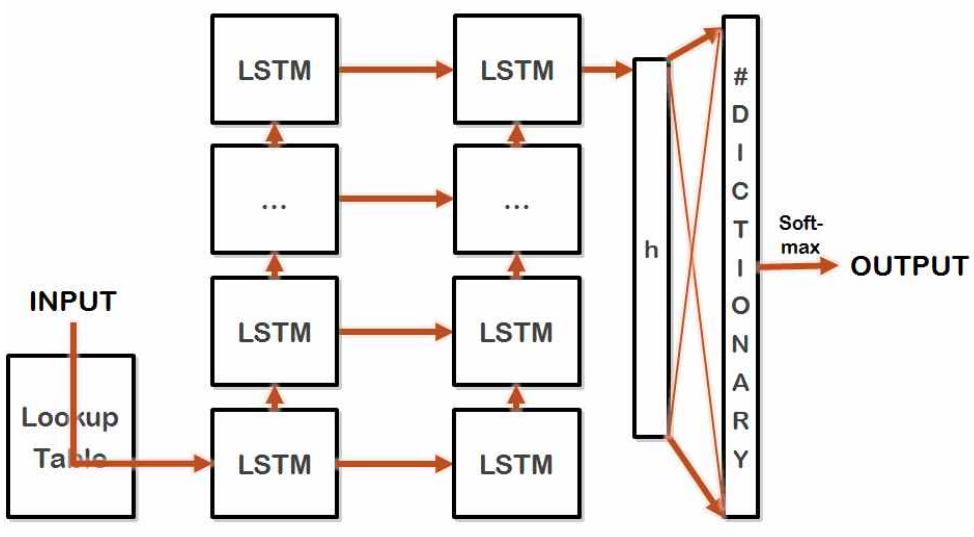



**[그림 6]** Augmented TextQA 구조

Augmented TextQA를 위한 방법은 크게 3가지로 나눌 수 있다 [그림 6]. (1) Stanford CoreNLP를 이용하여 지문 속에서 주요 정답이 되는 단어, 키워드를 추출하는 부분과, (2) 지문과 질문대신 답을 이용하여 질문을 생성하는 부분, (3) (2)에서 학습한 모델을 이용하여 키워드에 대한 질문 Q′을 생성하고, 이 생성된 Q′와 그 정답인 키워드를 기존 데이터에 추가하여 TextQA를 하는 Augmented TextQA 부분으로 나눌 수 있다. 각 모델의 구조들은 3.2.2에 기술한 모델 구조 중에서 실험을 통해 가장 성능이 좋은 구조를 선택하였다.

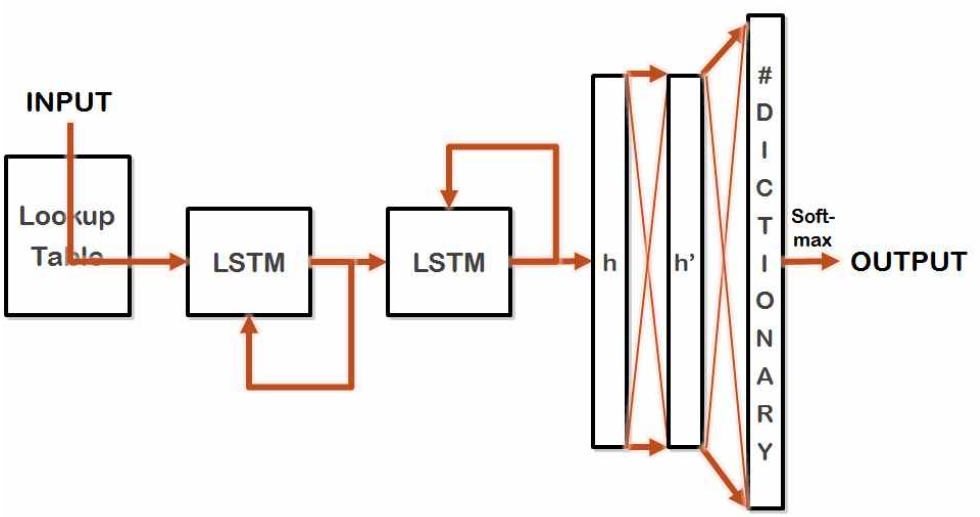
### 3.2.2 모델 구조

실험에는 n-layer Seq2Seq, Conv + Seq2Seq, Seq2Seq + 2-Fully Connected Layer 모델을 활용해보았다. 각 모델을 도식화하면 [그림 7, 8, 9]와 같다.

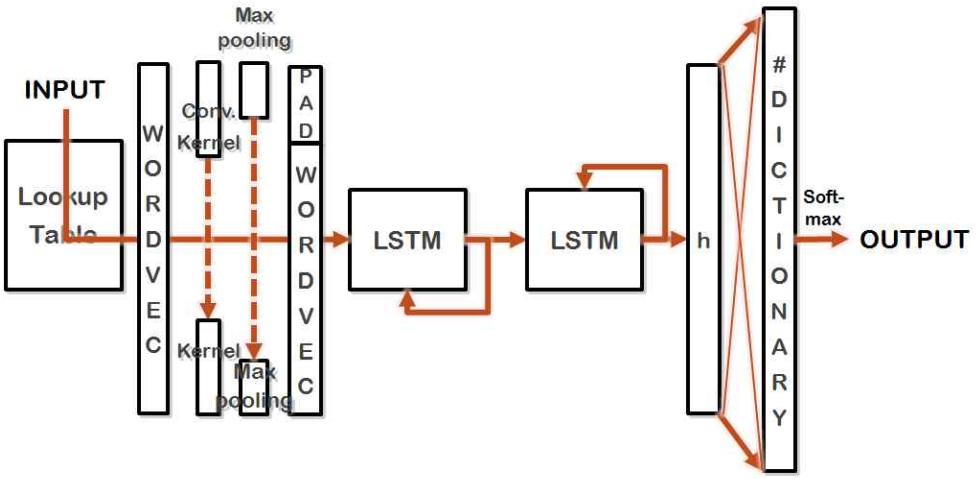



[그림 7] n-layer Seq2Seq

n-layer Seq2Seq 모델 [그림 7]은 기존 Seq2Seq 모델에서 RNN 계열의 인공신경망인 LSTM (Long-Short Term Memory) (Hochreiter & Schmidhuber, 1997)을 여러 층으로 쌓아올려 매 시간 축마다 인코딩된 입력과, 디코딩 되는 출력이 각각 dependency를 유지하도록 설계한 모델이다. 마지막 LSTM의 hidden vector(h)를 1층짜리 fully connected layer를 이용하여 단어 사전에 있는 단어의 수만큼의 output vector로 표현한 후 Softmax 함수를 거쳐 매 time step 마다 출력 단어를 생성한다.




[그림 8] Seq2Seq + 2FCL




[그림 9] Conv + Seq2Seq

Seq2Seq + 2-Fully Connected Layer [그림 8]는 기존 Seq2Seq 모델에서 마지막 fully connected layer를 두 층으로 쌓은 모델이다. Seq2Seq의 결과로 생성된 hidden vector (h)를 추가된 fully connected layer를

이용하여 더 추상적인 결과를 hidden vector (h′)로 얻은 뒤, 마찬가지로 단어 사전 속 단어 수만큼의 output vector로 표현하여 Softmax 함수를 통해 매 time step 마다 출력 단어를 생성한다.

Conv + Seq2Seq는 입력에 CNN (Convolutional Neural Networks) (LeCun & Bengio, 1995)을 돌려 나온 결과를 Seq2Seq의 입력으로 하여 최종 LSTM의 hidden vector를 얻는다. 이후 출력은 앞서 설명한 방법들과 동일하게 얻는다. CNN의 convolution layer는 이미지를 처리할 때와는 다르게 1차원 커널을 사용하며 1차원 maxpooling을 한 뒤 Seq2Seq의 입력 차원에 맞도록 zero-padding을 한다. 단순히 입력을 Seq2Seq에 넣기보다는 convolution layer를 통해 지문으로부터 적절한 feature를 추출하여 학습하기를 기대하였다.

3.2.1의 전체 구조에 따라 우리는 질문을 생성하는 모델과 TextQA 모델을 만들어야했다. 따라서 위에 제시한 세 가지 모델 구조 가운데 질문 생성 모델과 TextQA 모델에 적합한 모델 구조를 찾기 위해 각 모델의 성능을 테스트했다.

# IV. 실 험

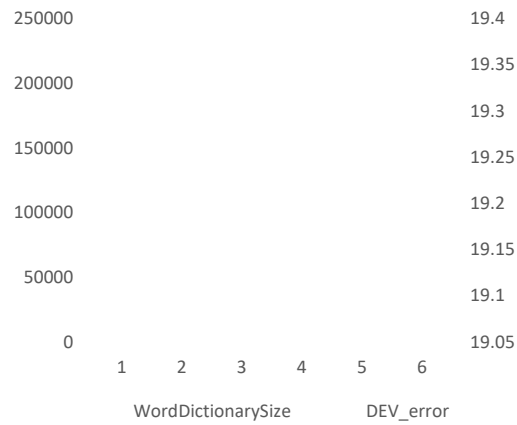
## 4.1. 전처리

SQuAD 데이터뿐만 아니라 거의 모든 자연어 데이터들은 정형화되어 있지 않기 때문에 전처리 과정이 필요하다. 그 과정에서 어느 형태의 단어까지 의미단위로 생각할지 판단하는 tokenize 방법 또는 얼마만큼의 단어의 양을 커버하도록 할지, 다시 말해 corpus 크기를 정의하는 방법 등에 의해 성능이 달라진다. 단순히 모든 단어와 특수문자를 처리하기에는 단어 사전의 크기가 커지게 되어서 계산 량이 기하급수적으로 증가할 수 있다. 따라서 매우 드물게 사용되는 단어를 out-of-vocabulary 심볼을 활용하여 한 토큰으로 처리하여 계산량을 줄이고 학습셋에 존재하지 않아 단어 사전에 구축하지 못한 단어도 처리할 수 있도록 한다. 각 데이터마다 최적의 tokenize 방법 다르기에, SQuAD에서 다양한 방법의 실험을 해보았다: (1) 토큰을 띄어쓰기 단위로 설정, (2) (1)의 방법에 전부 소문자화, (3) 토큰을 특수 기호를 제외한 단어로 설정, (4) (3)의 방법에 전부 소문자화. 그러나 (1) 의 방법은 단어의 기본형에 특수기호가 붙은 형태를 하나의 새로운 토큰으로 인식하는 문제점이 있다. 예를 들어 A. 와 A를 다르게 인식하여 A를 학습할 수 있는 기회는 줄어들고 단어 사전의 크기만 증가될 것이다. 반면 (3)은 영어에서 흔히 쓰이는 축약형, out-of-date 같은 특수 기호를 포함한 단어들을 처리하기에 적절하지 않았다. 또한 (2), (4)에서는 token을 모두 소문자화 하였는데, 이는 문장에서 대문자로 쓰인 단어들을 같은 소문자 token으로 묶여줄 수 있지만 이는 반드시 대문자로 써야하는 대명사와 같은 정보를 잃어버릴 수 있다.

따라서 각각의 tradeoff를 적절히 조절할 수 있는 tokenize 방법이 필요했다. 우리는 적절한 tokenize 방법이 활용되었을 때 기본적인 TextQA 모델에서 가장 성능이 잘 나올 것이라 가정하고, 1-layer Seq2Seq 모델에서 TextQA 성능 변화를 살펴보며 tokenize를 위한 정규표현식을 튜닝하였다. 최종적으로 얻은 정규표현식은

```
^[%(]?[%"\'%“£€]?(%$?[a-zA-Z0-9&%.,äêéöõüÄÖÜßŻ]*[a-zA-Z0-9%-&äêéöõüÄÖÜßŻ]+%%?)
```

과 같으며, 이 정규표현식을 사용하였을 때를 (5), (5)에서 소문자화를 추가한 방법을 (6)이라고 할 때, 각 방법들의 토큰의 크기와 TextQA 성능은 [그림 10]로 나타낼 수 있었다.




**[그림 10]** 전처리 방법에 따른 단어 사전 크기와 각 성능

실험 결과, 단어 사전의 크기는 (1)>(2)>(5)>(6)>(3)>(4)로 나타났으나 성능은 (5)에서 가장 좋았다. 위 실험을 통해 단순히 많은 단어를 커버 가능한 tokenize 방법이 언어 모델에 좋은 것은 아니라는 것을 확인했으며, 최적의 성능을 보인 (5)의 tokenize 방법을 이후 실험에 계속 활용하기로 하였다.

## 4.2. TextQA

3.2.2에 기술했던 Seq2Seq 기반 모델들을 이용해 TextQA를 수행해보았다. [그림 11]은 각 모델의 학습 시간에 따른 validation 데이터에서의 perplexity를 표현한 그래프이다.


**[그림 11]** TextQA 모델 성능

Perplexity는 모델이 다음에 생성할 단어를 얼마나 잘 예측했는지에 대한 척도로, 주어진 파라미터 에 대한 단어의 Negative log-likelihood (NLL)를 계산하여 평균 NLL의 exponential을 취한 값을 활용한다.

$$LL = - \sum_{=1} \ln P(\mathrm{word}_i | \theta)$$

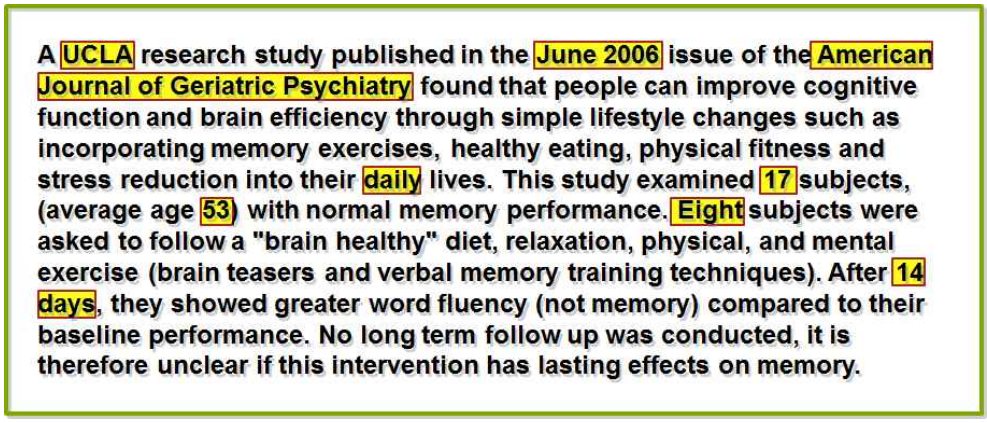
$$rplexity = \exp\left(\frac{NLL}{Number\ of\ tokens}\right)$$

따라서 Perplexity가 낮을수록 모델이 해당 단어를 생성할 확률이 높다는 것을 의미한다.

실험에 사용된 Seq2Seq의 hidden layer 크기는 128로 설정하였으며, convolution layer의 커널 크기는 4, pooling 크기는 2일 때 가장 성능이 좋았다. TextQA 실험에서는 Conv+Seq2Seq 모델의 성능이 가장 좋았으며, 그 다음은 2-layer Seq2Seq, 3-layer Seq2Seq, Seq2Seq+2FCL 순으로 성능이 좋았다. n-layer Seq2Seq 중에서는 2-layer, 3-layer가 1-layer, 4-layer보다 좋은 것으로 성능이 나타났다. 1-layer Seq2Seq보다 조금 더 복잡한 모델인 Seq2Seq+2FCL이 성능이 더 뛰어난 것으로 보아 1-layer에서는 underfitting이, 반대로 4-layer에서는 모델을 계속해서 쌓았으나 validation 데이터를 제대로 예측하지 못하는 overfitting이 발생한 것으로 추측할 수 있었다.

## 4.3. 키워드 추출

Stanford CoreNLP (Manning et al., 2014)를 사용하여 각 지문에서 답이 될 수 있는 키워드들을 추출하였다. 최대한 키워드 추출 오류를 줄이기 위해서 고유명사, 년도, 시간, 금액 및 숫자, 장소와 같은 명백하게 답이 될 수 있는 키워드를 위주로 추출하였다. 그 결과 학습 데이터에서 총 244,808개의 키워드들을 추출하였으며 각 키워드는 1개 단어 이상의 문장으로 구성되었다. [그림 12]는 주제가 Memory인 지문에서 추출된 키워드의 예시이다.

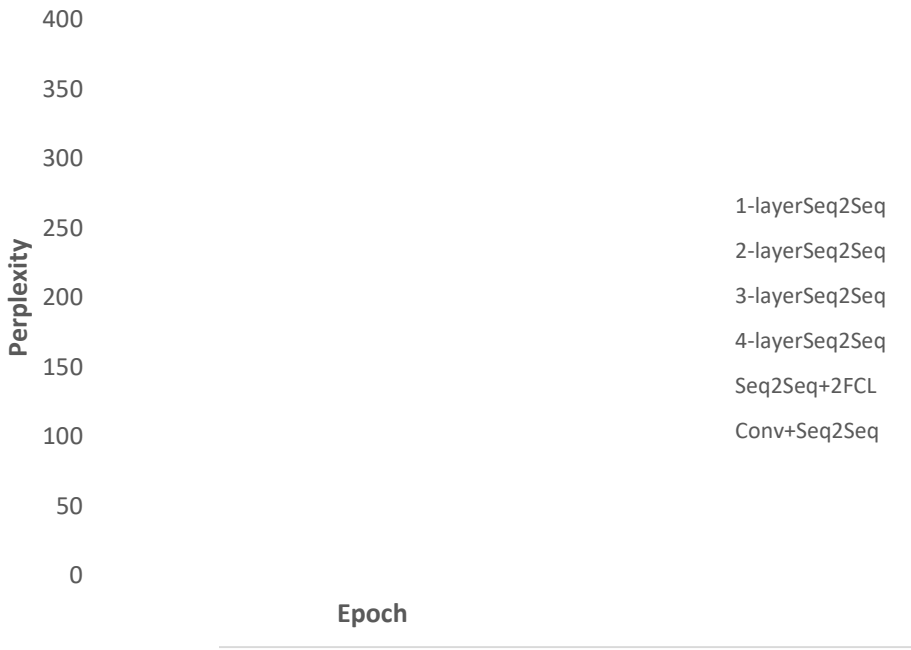



**[그림 12]** Stanford CoreNLP를 이용한 키워드 추출 예시

[그림 12]를 살펴보면 고유명사(예: UCLA)와 날짜, 수량 (예: June 2006, 17) 등의 키워드를 잘 추출하는 것을 확인할 수 있었다. 이 자잘한 키워드들은 기존 QA데이터에서 포함하고 있지 않은 내용이었으며, 이러한 키워드까지 잘 학습한다면 모델의 성능이 더 좋아질 것으로 기대하였다.

## 4.4. 질문 생성

앞서 추출한 키워드를 활용하기 위해 주어진 데이터를 이용하여 질문을 생성하는 모델을 학습하였다. [그림 13]은 validation 데이터에 대한 질문 생성 모델의 성능을 나타낸다.




**[그림 13]** 질문 생성 모델 성능

전반적으로 질문 생성이 QA보다 훨씬 복잡한 문제였다. 결과는 마찬가지로 Conv+Seq2Seq 모델이 가장 좋은 성능을 보였다. n-layer Seq2Seq 모델은 2-layer, 1-layer, 3-layer, 4-layer 순으로 성능이 좋았으며 Seq2Seq+2FCL은 2-layer와 1-layer의 중간 수준의 성능을 보였다. 이 결과를 통해 질문 생성 문제는 지문 속에서 적절한 feature들을 이용

하여 질문을 생성하되 1-layer보다는 2-layer 수준의 복잡한 모델이 필요하다는 것을 알 수 있었다. 따라서 우리는 이후 실험에서 질문을 생성할 때 가정 성능이 좋았던 Conv+Seq2Seq 모델을 이용하였다. 이 실험 결과의 예시는 부록 2에 첨부하도록 한다.

## 4.5. Augmented TextQA

Conv+Seq2Seq로 만든 질문 생성 모델과 4.3에서 추출한 키워드를 이용하여 Augmented TextQA를 수행하였다. 학습 데이터는 추출된 키워드의 개수와 동일하게 87,599에서 332,407로 약 4배가량 증가하였으며 같은 지문을 추가된 키워드로 인해 여러 번 학습하다보니 학습데이터에 대한 perplexity가 빠르게 감소하는 것을 확인할 수 있었다. [그림 14]는 validation 데이터에 대한 Augmented TextQA 모델의 성능이다.


**[그림 14]** Augmented TextQA 모델 성능

Conv+Seq2Seq, 2-layer Seq2Seq, Seq2Seq+2FCL의 순서로 성능이 좋았다. 실험 결과의 예시는 부록 3에 첨부하도록 한다.

# V. 결 과

TextQA 모델과 Augmented TextQA를 수행하였을 때의 성능을 비교하면 [그림 15]와 같다.


**[그림 15]** TextQA vs. Augmented TextQA 모델 성능 비교

모든 모델에서 Augmented TextQA 방법을 사용할 때 성능 향상이 있었다. Augmented TextQA의 성능은 순서대로 Conv+Seq2Seq, 2-layer Seq2Seq, Seq2Seq+2FCL 순으로 나타났다. 세 가지 모델에서 약 11%의 성능 향상이 있었으며 그 결과 가장 성능이 좋지 않았던 Seq2Seq+2FCL이 Augmented TextQA로 인해 Augmented 되지 않은 2-layer Seq2Seq보다 좋은 성능을 보이기도 했다.

# VI. 결론 및 논의

본 논문은 주어진 도메인 상에서 외부 텍스트 데이터의 추가 사용 없이 TextQA의 성능을 향상시키기 위해 데이터 증강 기법을 활용한 Augmented TextQA 방법을 제안한다. 그 결과 모델이 주어진 데이터뿐만 아니라 지문 내에서 답이 될 수 있는 주요 키워드를 예상하고 학습하여 더 많은 질문에 올바르게 대답할 수 있는 모델을 만들 수 있었다. Augmented TextQA 방법은 학습 할 때 키워드를 보고 그 예상 문제를 생각하며 공부하는 사람의 학습 방법과 유사하며, Bottou (2014)가 주장했던 키워드 추출, 예상 질문 생성, 추가 학습의 문제 해결 과정을 모듈을 쌓아 구현하는 Reasoning의 과정을 따르고 있다.

Augmented TextQA는 기존 TextQA에 비해 질문을 생성하는 모델을 추가로 만들어야한다는 단점이 있지만, 성능의 소수점 차이에 의해서 등수가 변화하는 TextQA 대회에서는 충분히 시도할 가치가 있는 방법이라고 생각된다. 그 다음으로 질문을 생성하는 모델을 학습하는 과정이나 키워드를 추출하는 과정에서 오류가 발생하겠지만 두 오류를 충분히 최소화시킨다면 TextQA 모델이 기존 데이터에 overfitting하지 않도록 모델을 일반화(generalization)하는 역할을 해줄 수 있을 것이라 기대한다.

TextQA에 대한 추가 연구로 앞서 제시했던 Seq2Seq 기반 모델 3개를 적절히 혼합하여 성능이 더 뛰어난 모델을 만들고, SQuAD의 특성에 걸맞도록 모델이 생성한 답이 지문 속 어디에 해당하는지를 찾는 추가적인 엔지니어링 과정을 통해 Exact Match와 F1 score를 향상시켜 리더보드에 등록해보았으면 한다.

# 참고문헌


Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(Aug), 2493-2537.

Chai, K. M. A., Chieu, H. L., & Ng, H. T. (2002, August). Bayesian online classifiers for text classification and filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 97-104). ACM.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics, 18*(4), 467-479.

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology, 6*(3), 361-365.

Rieger, B. B. (1991). On distributed representation in word semantics. Berkeley, CA: International Computer Science Institute.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016, June). Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning* (pp. 1378-1387).

Wang, W. Yang, N. Wei, F. Chang, B. & Zhou, M. (2017). Gated selfmatching networks for reading comprehension and question answering. *in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (ICML-14) (pp. 1188-1196).

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global Vectors for Word Representation. In *EMNLP (Vol. 14*, pp. 1532-1543).

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014, November). Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 157-166). ACM.

Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. arXiv preprint arXiv:1502.01710.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Luong, M. T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. arXiv preprint arXiv:1511.06114.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016, March). End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing* (ICASSP), 2016 IEEE International Conference on (pp. 4945-4949). IEEE.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G. P., & Vanderwende, L. (2017). Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. arXiv preprint arXiv:1701.08251.

Dong, L., & Lapata, M. (2016). Language to logical form with neural

attention. arXiv preprint arXiv:1601.01280.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)* (pp. 55-60).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

Bottou, L. (2014). From machine learning to machine reasoning. *Machine learning, 94*(2), 133-149.

## ABSTRACT

# Data Augmentation Technique with Knowledge Extraction for Text Question Answering by Deep Neural Networks

**Hwiyeol Jo**

School of Computer Science & Engineering

The Graduate School

Seoul National University

Teaching a machine that communicates with people in natural language is a dream of artificial intelligence researchers. However, due to the numerous exceptions and uncertainties in language, language modeling with traditional rule-based approach has several limitations. Recently, language modeling research has been continuing to overcome the limitation by introducing deep learning algorithms. Among the studies of natural language processing, TextQA is a study that generates an answer by looking at the context and

question, and it is suitable to test how well the language model understands the natural words. Despite many studies, it has not yet reached the performance of people. In this paper, we propose Augmented TextQA for improving the performance of TextQA models. Augmented TextQA first creates a question generation model using an answer and a context, and generates a question through the model by inputting a keyword extracted from the context. Finally, we try to improve the performance of TextQA model by adding the augmented data. Experimental results using the SQuAD (Stanford Question Answering Dataset) show that the model performance is improved in Augmented TextQA compared with original TextQA. Furthermore, using Augmented TextQA can augment the data on the domain of train data without using external data, so it would learn the representation of the word in a specific domain more appropriately. This method also can generalize the TextQA model not to be limited to the original train data but to fit on the extended train data.

**Keywords: Data Augmentation, Deep Learning, Natural Language Processing, Text Question-Answering**

**Student Number: 2015-22915**

# 부 록

**부록 1.** TextQA 결과 예시

**부록 2.** 질문 생성 결과 예시

**부록 3.** Augmented TextQA 결과 예시

## [부록 1] TextQA 결과 예시 (좋음 - 보통 - 나쁨 순)

Q : The term imperialism has been applied to western countries and which eastern county

Imperialism is a type of advocacy of empire Its name originated from the Latin word imperium which means to rule over large territories Imperialism is a policy of extending a country power and influence through colonization use of military force or other means Imperialism has greatly shaped the contemporary world It has also allowed for the rapid spread of technologies and ideas The term imperialism has been applied to Western and Japanese political and economic dominance especially in Asia and Africa in the 19th and 20th centuries Its precise meaning continues to be debated by scholars Some writers such as Edward Said use the term more broadly to describe any system of domination and subordination organised with an imperial center and a periphery

Ground Truth : Japan

Model Answer : Japan

Q : What town in upstate New York was settled by Huguenots

Huguenot immigrants did not disperse or settle in different parts of the country but rather formed three societies or congregations one in the city of New York another 21 miles north of New York in a town which they named New Rochelle and a third further upstate in New York The Huguenot Street Historic District in New York has been designated a National Historic Landmark site and contains the oldest street in the United States of America A small group of Huguenots also settled on the south shore of Staten Island along the New York Harbor for which the current neighborhood of Huguenot was named

Ground Truth : New York

Model Answer : the York

Q : Along with the American Institute of Electrical Engineers what other institute eventually became the IEEE

Tesla served as a vice president of the American Institute of Electrical Engineers the forerunner along with the Institute of Radio Engineers of the modern-day IEEE from 1892 to 1894

Ground Truth : the Institute of Radio Engineers

Model Answer : the Middle of the

[부록 2] 질문 생성 결과 예시 (좋음 - 보통 - 나쁨 순)

Answer : grace of God which sustains the believers in the journey toward Christian Perfection

<unk> Grace is that grace of God which sustains the believers in the journey toward Christian Perfection a genuine love of God with heart soul mind and strength and a genuine love of our neighbors as ourselves <unk> grace enables us to respond to God by leading a <unk> and Christ-like life aimed toward love Wesley never claimed this state of perfection for himself but instead insisted the attainment of perfection was possible for all Christians Here the English Reformer parted company with both Luther and Calvin who denied that a man would ever reach a state in this life in which he could not fall into sin Such a man can lose all inclination to evil and can gain perfection in this life

Ground Truth : What is <unk> Grace

Model Question : What is <unk> Grace

Answer : 1884

Tesla gained experience in telephony and electrical engineering before emigrating to the United States in 1884 to work for Thomas Edison in New York City He soon struck out on his own with financial backers setting up laboratories and companies to develop a range of electrical devices His patented AC induction motor and transformer were licensed by George Westinghouse who also hired Tesla for a short time as a consultant His work in the formative years of electric power development was involved in a corporate alternating current current War of Currents as well as various patent battles

Ground Truth : In what year did Nikola Tesla emigrate to the United States

Model Question : What what year did

Answer : Denver

Denver took the opening kickoff and started out strong with <unk> <unk> completing an <unk> pass to tight end Owen Daniels and a <unk> throw to receiver Andre Caldwell A pair of carries by C J Anderson moved the ball up 20 yards to the Panthers <unk> line but Carolina defense dug in over the next three plays First linebacker <unk> Thompson tackled Ronnie <unk> for a <unk> loss Then after an <unk> Thomas Davis tackled Anderson for a <unk> gain on third down forcing Denver to settle for a 3 <unk> lead on a Brandon <unk> <unk> field goal The score marked the first time in the entire postseason that Carolina was facing a deficit

Ground Truth : Which team has the first kickoff

Model Question : What two used year

[부록 3] Augmented TextQA 결과 예시 (좋음 - 보통 - 나쁨 순)

Q : Since this was the 50th one what was the theme color for the Super Bowl

Various <unk> promotions and initiatives were held throughout the 2015 NFL season to tie into the Golden Super Bowl <unk> logos were implemented across the NFL properties and painted on fields the numbering of the 50-yard line on fields was colored gold and beginning on week 7 all sideline jackets and hats featured <unk> logos Gold <unk> were given to each high school that has had a player or coach appear in the Super Bowl and homecoming events were also held by Super <unk> teams at games <EOS>

Ground Truth : gold

Model Answer : gold

Q : When were the elections that produced a higher vote by Protestant Nazi sympathizers than by Catholics

Since the 1980s Lutheran Church denominations have repudiated Martin Luther statements against the Jews and have rejected the use of them to incite hatred against Lutherans <unk> et al 1970 survey of 4,745 North American Lutherans aged 15 <unk> found that compared to the other minority groups under consideration Lutherans were the least <unk> toward Jews Nevertheless Professor Richard Dick Geary former Professor of Modern History at the University of Nottingham England and the author of Hitler and Nazism Routledge 1993 wrote in the journal History Today an article on who voted for the Nazis in elections held from <unk> where he claimed that from his research he found that the Nazis gained disproportionately more votes from Protestant than Catholic areas of Germany

Ground Truth : <unk>

Model Answer : <unk>

Q : Who challenged the plague theory first

The plague theory was first significantly challenged by the work of British bacteriologist J F D Shrewsbury in 1970 who noted that the reported rates of mortality in rural areas during the 14th-century pandemic were inconsistent with the modern bubonic plague leading him to conclude that contemporary accounts were <unk> In 1984 zoologist Graham <unk> produced the first major work to challenge the bubonic plague theory directly and his doubts about the identity of the Black Death have been taken up by a number of authors including Samuel K Cohn Jr 2002 David <unk> 1997 and Susan Scott and Christopher Duncan 2001 <EOS>

Ground Truth : British bacteriologist J F D Shrewsbury

Model Answer : Kathmandu and Kennedy